DOCUMENT RESUME

ED 302 567 TM 012 569

AUTHOR Cashin, William E.
TITLE Student Ratings of Teaching: A Summary of the Research. IDEA Paper No. 20.
INSTITUTION Kansas State Univ., Manhattan. Div. of Continuing Education.
PUB DATE Sep 88
NOTE 7p.
AVAILABLE FROM Center for Faculty Evaluation and Development, Division of Continuing Education, Kansas State University, Manhattan, KS ($1.00).
PUB TYPE Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Bias; College Faculty; *College Students; *Educational Research; Evaluation Criteria; Factor Analysis; Higher Education; Literature Reviews; *Professors; Student Attitudes; *Student Evaluation of Teacher Performance; *Teacher Effectiveness; Teacher Evaluation
IDENTIFIERS Rater Reliability

ABSTRACT

Conclusions of major reviews of the literature on student ratings of college teachers are summarized for the period 1971 to September 1988. Several factor analytic studies have concluded that student ratings are multidimensional and that no single item will be useful for all purposes. Consistency, stability, and generalizability are aspects of reliability generally discussed. The validity of student rating forms has been defined in terms of student learning, through the self-ratings of the instructor, and through the ratings of others. In spite of widespread faculty concern, research has shown relatively few sources of bias in student ratings of teachers. Some relationships that may bias student rating data include: (1) faculty rank; (2) expressiveness; (3) student motivation; (4) expected grades; (5) course level; (6) academic field; (7) workload or difficulty; (8) ratings that are not anonymous; (9) presence of the instructor as the rating is made; and (10) purpose of the ratings. In general, student ratings have been found reliable, valid, and relatively free of bias. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED302567

IDEA Paper No. 20

center for FACULTY EVALUATION & DEVELOPMENT
Division of Continuing Education
Kansas State University

September 1988

Student Ratings of Teaching: A Summary of the Research

William E. Cashin
Kansas State University

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

WILLIAM CASHIN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

"... class average student ratings are (a) multidimensional, (b) reliable and stable, (c) primarily a function of the instructor who teaches a course rather than the course that is taught, (d) relatively valid against a variety of indicators of effective teaching, (e) relatively unaffected by a variety of variables hypothesized as potential biases, and (f) seen to be useful by faculty as feedback about their teaching, by students for use in course selection, and by administrators for use in personnel decisions."

(Marsh, 1984, p. 707)

There are more than 1,300 articles and books dealing with research on student ratings of teaching. The last IDEA Paper on the subject, entitled **Reliability, Validity, and Generalizability of Student Ratings of Instruction**, was written seven years ago (Aubrecht, 1981). No major study published since then has challenged that paper's conclusions, but a number of excellent reviews of the literature provide further support. This paper will attempt to summarize the conclusions of the major reviews of the student rating literature from Costin, Greenough, and Menges (1971) to the present. That literature is extensive and complex. Obviously, a paper of six pages can offer only broad, general conclusions. I strongly recommend that the interested reader consult the various reviews and their individual references for details.

The ERIC descriptor for student ratings is "student evaluation of teacher performance." I suggest that it is much more useful to think of these student reactions (and peer "evaluations," etc.) as data that need to be interpreted. Using the term "student ratings" rather than "student evaluations" helps to distinguish between the people who provide the information (sources of data) and the people who interpret it (evaluators).

Viewing student ratings as data rather than as evaluations may also help to put them in proper prospective. Writers about faculty evaluation are almost universal in recommending the use of **multiple** sources of data. No single source of data, including student rating data, provides sufficient information to make a valid judgment about teaching effectiveness.

Multidimensionality

There have been a number of factor analytic studies (see Cohen, 1981; Feldman, 1976b; and Kulik & McKeachie, 1975, for references) which conclude that student rating forms are multidimensional, i.e., that they measure **several aspects** of teaching. Put another way, no single student rating item nor set of related items will be useful for all purposes.

Marsh's (1984) SEEQ (Students' Evaluations of Educational Quality) form has nine dimensions: learning/value, enthusiasm, organization, group interaction, individual rapport, breadth of coverage, exams/grades, assignments, and workload. Other student rating forms have items measuring some or all of those dimensions. In several of his reviews of the literature, Feldman (1976b, 1983, 1984, and 1987) categorized student ratings items into as many as 19 different dimensions. In his most recent review (Feldman, in press) he uses 22 dimensions. Therefore, when interpreting student rating data, we must **distinguish among the various items and their dimensions**. Averaging all of the items is **not** appropriate. And it is important to remember that no single rating item can serve all purposes.

Reliability

In the educational measurement literature, reliability covers consistency, stability, and generalizability. For student rating items, reliability is usually concerned with **consistency**, with interrater agreement, which varies depending upon the number of raters, i.e., the more raters, the more reliable. For example, with the IDEA system (Cashin & Perrin, 1978), the average item reliabilities are:

10 raters = .69;
20 raters = .81;
40 raters = .89.

Similar or higher reliabilities are typically found with other well-designed forms, i.e., forms developed with the assistance of someone knowledgeable about educational measurement. It is generally suggested that items with fewer than ten raters (reliabilities below .70) be interpreted with **particular caution**.

Stability is concerned with agreement between raters overtime. A longitudinal study (see Marsh & Overall, 1979, and Overall & Marsh, 1980) compared end-of-course ratings with ratings by the same students years later (at least one year after graduation). The average correlation was .83.

BEST COPY AVAILABLE

TM012569

Generalizability is concerned with how confident we can be that our data accurately reflect the instructor's teaching effectiveness in general, not just in this particular course this term. A study conducted by Marsh (1982) illustrates the question. He studied data from 1,364 courses, dividing them into four categories: the same instructor teaching the same course but in different terms; the same instructor teaching a different course; different instructors teaching the same course; and different instructors teaching different courses. This permitted him to study the differential effects of the instructor and of the course. He then correlated student ratings in the four different categories, separating items related to the instructor (e.g., enthusiasm, organization, discussion) from background items (e.g., student's reason for taking the course, workload). The average correlations are given below, with those for background items in parentheses.

	Same Course	Different Course
Same Instructor	.71 (.69)	.52 (.34)
Different Instructor	.14 (.49)	.06 (.21)

The instructor related correlations were higher for the same instructor, even when teaching a different course. The correlations for the background items—more tied to the course than the instructor—were higher for the same course. Marsh concluded that the instructor, not the course, is the primary determinant of the student rating items. Marsh's results are reasonably comparable to other generalizability studies (Bausell, Schwartz, & Purohit, 1975; Gillmore, Kane, & Naccarato, 1978; and Hogan, 1973).

When making personnel decisions we want to use the data to make judgments about the instructor's general teaching effectiveness. When considering student rating (remembering that we need other kinds of information beyond student ratings), the following seem reasonable rules of thumb: If the instructor teaches only one course (e.g., part-time instructors) **consistent** ratings from two different terms may be sufficient. For most instructors, however, I would recommend ratings from two or more courses from every term for at least two years. If there are fewer than 15 raters in the classes, data from additional classes are recommended.

Validity in General

In educational measurement, the basic question concerning validity is: does the test measure what it is supposed to measure? For student ratings this translates into: to what extent do student rating items measure some aspect of teaching effectiveness? Unfortunately, there is no agreed upon definition of "effective teaching" nor any single, all-embracing criterion. The best that one can do is examine data that either support or challenge the conclusion that student ratings reflect effective teaching.

Validity—Student Learning

Theoretically, the best criterion of effective teaching is student learning. In general, the students of more effective teachers should learn more. A number of studies have attempted to study this hypothesis by comparing multiple-section courses. In the typical study, different instructors taught different sections of the same course, using the same syllabus and textbook, and most important, using a common external final exam, i.e., an exam developed by someone **other** than the instructors. Cohen (1981, 1986) reviewed these studies. Using the students' grades on the external exam as the measure of student learning, he examined correlations between the exam grade and various student rating items. The average correlations from Cohen (1981) are given below.

With ratings of their **achievement** (how much students thought they learned)	.47
With ratings of **overall course effectiveness**	.47
With ratings of **overall instructor effectiveness**	.44
With ratings on **teacher skill** dimension (e.g., explains clearly)	.50
With ratings on **teacher structure** dimension (e.g., uses class time well)	.47
With ratings on **teacher rapport** dimension (e.g., is friendly)	.31
With ratings on **teacher interaction** dimension (e.g., facilitates classroom discussion)	.22

Note on Interpreting Validity Correlations: Earlier I suggested that reliability correlations **below** .70 were undesirable. However, in the social sciences validity correlations **above** .70 are unusual, especially when studying complex phenomena, such as student learning. As a rule of thumb, I suggest that student rating validity correlations between .00 and .19, even when statistically significant, are not usually practically useful. Correlations between .20 and .49 are practically useful. Correlations between .50 and .70 are very useful but they are rare when studying complex phenomenon.

Using the above rule of thumb, the average correlations reported by Cohen (1981) are generally useful. These relationships tend to support the validity of student ratings because the classes in which the students gave the instructor higher ratings tend to be the classes where the students learned more, i.e., scored higher on the external exam. On the other hand, the correlations are far from perfect because many of the variables which relate to students' learning will be related to student characteristics (e.g., ability and motivation), not to instructor characteristics.

Validity—Instructor's Self Ratings

Researchers have sought for a criterion of effective teaching that would be acceptable to faculty. One possibility is the self ratings of the instructor himself or herself. In a review of the literature, Marsh (1984) cites ten studies which correlated instructor's self ratings with student ratings. The correlations varied from .20 to .69, averaging .41. In one study (Marsh, Overall, & Kesler, 1979) instructors were asked to rate two of their courses in order to see if the course that the instructor rated higher was also rated higher by the students. The median correlation between the instructor's self ratings and the students' ratings was .49. Such studies provide further support for the validity of the students' ratings.

Validity—The Ratings of Others

If one is willing to grant that the ratings of administrators, colleagues, and alumni have some validity, then student ratings share that validity.

Student ratings correlate with **administrators ratings**, ranging from .47 to .62 (Kulik & McKeachie, 1975).

Student ratings correlate with **colleague's ratings**, .48 to .69 (Kulik & McKeachie, 1975).

Student ratings correlate with **alumni ratings**, .40 to .75 (Overall & Marsh, 1980; Braskamp, Brandenberg, & Ory, 1984). This belies the conventional wisdom that the students will come to appreciate our teaching after they get into the real world.

Validity—Possible Sources of Bias—The Good News

One need not talk with faculty very long to be aware of their concern about possible biases in student ratings. Some have suggested that bias be defined as anything **not under the control of the instructor**. Marsh (1984) argues against this definition. For example, grading leniency—instructors giving higher grades than the students earned—would not be considered a bias using this definition. Marsh suggests that student ratings are biased only to the extent that they are influenced by variables **not related to teaching effectiveness**. When using ratings for personnel decisions or the instructor's improvement, I would suggest an even narrower definition, restricting bias to variables **not** a function of the **instructor's** teaching effectiveness. Student motivation or class size may impact teaching effectiveness, but instructors should not be faulted if they are less effective teaching large classes of unmotivated students than their colleagues are with small classes of motivated students. In this case, student motivation and class size, although they are related to teaching effectiveness, are not a function of the instructor's characteristics, but of student and course characteristics, and so they should be considered sources of bias. They should, however, be controlled for by using appropriate comparative data (Cashin, 1985).

Despite widespread faculty concern, the research has uncovered relatively few sources of bias. Generally the following variables tend to show **little or no** relationship to student ratings.

A. INSTRUCTOR VARIABLES **not** related to student ratings:

1. **Sex of the instructor**. But a few studies suggest there may be a sex of instructor by sex of student interaction, e.g., Basow and Distenfeld, 1985; Basow and Silberg, 1987.

2. **Instructor's age, and teaching experience**. However, when differences have been found, they tend to be negative; i.e., older faculty receive **lower** ratings (Feldman, 1983). (Few studies have tested for curvilinearity; ratings may tend to increase for the first ten or so years of experience then begin to decline.)

3. **Instructor's personality**, as measured by instructor's self report (personality inventories, self-description questionnaires, etc.) (Feldman, 1986). Only two of 14 traits showed practically significant correlations with overall evaluation items: positive self esteem ($r = .30$), and energy and enthusiasm ($r = .27$). (However, when studies used **students' and/or colleagues' perceptions** of the teacher's personality traits, 11 of the 14 traits had statistically significant average correlations with student ratings.)
Note: I would suggest that, even when agreed personal traits do correlate with student ratings, it might not be appropriate to control for them. These may be assets or deficits the instructor brings to his or her teaching and so should be taken into consideration.

4. **Research productivity**. The average correlation between research productivity and overall teaching effectiveness items is .12 (Feldman, 1987).

B. STUDENT VARIABLES **not** related to student ratings:

1. **Age of the student** (Menges, 1973).

2. **Sex of the student** (Bruskamp et al., 1984; Marsh, 1984).

3. **Level of the student**, e.g., freshman (McKeachie, 1979; Menges, 1973).

4. **Student's GPA** (Feldman, 1976a).

5. **Student's personality** (Marsh, 1984).

C. COURSE VARIABLES **not** related to student ratings:

1. **Class size**. Although there is a tendency for smaller classes to receive higher ratings, it is a very small inverse association (average $r = -.09$ (Feldman, 1984). The average correlation of class size for the 38 IDEA items is .18 (Cashin & Slawson, 1977).

2. **Time of day** (when the course is taught) (Menges, 1991; Feldman, 1978).

D. ADMINISTRATIVE VARIABLES **not** related to student ratings:

1. **Time during term** when ratings are collected (Aleamoni, 1981; Feldman, 1979).

Validity—Possible Sources of Bias—The Bad News

The research cited above suggests that many variables suspected of biasing student ratings are **not** correlated with them to any practically significant degree. For the following variables, however, the research suggests that there are relationships that may bias student rating data.

A. INSTRUCTOR VARIABLES **related** to student ratings:

1. **Faculty rank**. Regular faculty tend to receive higher ratings than graduate teaching assistants (Bruskamp et al., 1984). However, for the 38 IDEA items the average correlation with academic rank, including GTAs, is only .10 (Cashin & Slawson, 1977).

2. **Expressiveness**. The Dr. Fox effect (Naftulin, Ware, & Donnelly, 1973) suggests that student ratings may be more influenced by an instructor's style of presentation than by the substance of the content. In general, the literature (Marsh & Ware, 1982) suggests that, especially in studies involving an incentive and a test, manipulations of instructor expressiveness primarily influence items related to instructor enthusiasm, and manipulation of content coverage primarily influences items related to instructor knowledge and student exam performance. But making the subject interesting as well as informative helps students learn content, especially when incentives and testing are missing.

Since these variables, experience teaching, being expressive or interesting, and delivering content, **are related** to an instructor's teaching effectiveness, they are **not** sources of bias.

B. STUDENT VARIABLES **related** to student ratings:

1. **Student motivation**. Instructors are more likely to obtain higher ratings in classes where students had a prior interest in the subject matter (Marsh, 1984), or were taking the course as an elective (Aleamoni, 1981; Feldman, 1978). The average correlation of the IDEA motivation item, "I had a strong desire to take this course," with the other 37 items is .39.

Marsh (1987) concluded that **reason for taking the course**—which overlaps with student motivation—also is related to student ratings. Higher ratings were received from students who took a course for general interest, or as a major elective; lower ratings were received when the course is being taken as a major requirement or a general education requirement.

2. **Expected grades**. There tend to be positive, but low correlations (.10 to .30) between students ratings and expected grades (Feldman, 1976a; Howard & Maxwell, 1980 and 1982; Marsh, 1984). Three possible hypotheses have been proposed for these correlations:

Grading leniency—the instructor gives higher grades than the students deserve so the students give higher ratings than the instructor deserves.

Validity—the students who learned more earn higher grades and give higher ratings (therefore, student ratings are valid); and

Student characteristics—some student characteristics (e.g., high motivation) lead to both greater learning and, therefore, higher grades and higher ratings.

Marsh (1987, p. 421) concludes: "... while it is possible that a grading leniency effect may produce some bias in student ratings, support for this suggestion is weak and the size of such an effect is likely to be insubstantial in the actual use of student ratings." The two studies of Howard and Maxwell (1980 & 1982), which used IDEA data, provide support for a combination of the validity and student characteristics hypotheses.

C. COURSE VARIABLES related to student ratings:

1. **Level of the course**. Higher level courses, especially graduate courses, tend to receive higher ratings (Aleamoni, 1981; Braskamp et al., 1984). However, with the 38 IDEA items, course level only correlates .07 on average.

2. **Academic field**. Some studies suggest that humanities and arts type courses receive higher ratings than social science type courses, which, in turn, receive higher ratings than math-science type courses (Braskamp et al., 1984; Feldman, 1978; Marsh, 1984). This has been found for 44 academic fields with IDEA items (Cashin, Norma, & Hanna, 1987). Even partially controlling for differences in institutions, differences in ten academic fields accounted for 12% or more of the variance for half of the IDEA items (Cashin & Clegg, 1987). For a third of the items, differences in academic field accounted for 20% or more of the variance. **CAUTION**: Although there is increasing evidence that ratings for different fields differ, it is not clear why. If some fields are rated lower because they are more poorly taught, then these differences are not biases. On the other hand, if instructors in fields requiring more quantitative reasoning skills are rated lower because today's students are less competent in such skills—one of the hypotheses explaining why some fields are rated lower—then there is a bias which should be controlled for.

3. **Workload/difficulty**. These are correlated with student ratings (Marsh, 1984). However, contrary to faculty belief, they are correlated **positively**, i.e., students "give **higher** ratings in difficult courses where they have to work hard. Similar results are found with IDEA data (Cashin & Slawson, 1977). Average correlations with the remaining 37 IDEA items are:

Item 31 Amount of reading	.14
Item 32 Amount of other (non-reading) assignments	.11
Item 33 Difficulty of subject matter	.22
Item 35 I worked harder on this course than on most courses I have taken	.29

These results support the validity of student ratings rather than suggesting that they are biased.

D. ADMINISTRATIVE RATINGS related to student ratings:

1. **Non-anonymous ratings**. Signed ratings tend to be higher (Braskamp, et al., 1984; Feldman, 1979; Marsh, 1984). The hypothesis is that requiring students to sign their names inflates the ratings because some students are concerned about possible reprisals.

2. **Instructor present while students complete ratings**. These tend to be higher (Feldman, 1979; Marsh, 1984), possibly for the same reason as non-anonymous ratings.

3. **Purpose of the ratings**. Some studies have found that if the directions say the ratings will be used for personnel decisions, the ratings tend to be higher than if only used only by the instructor for improvement (Braskamp et al., 1984; Feldman, 1979; Marsh, 1984). Speculation is that the students tend to be lenient if the data will be used by someone other than the instructor.

Usefulness of Student Ratings

Although Marsh (1984) concluded that student ratings are "seen as useful by faculty as feedback about their teaching, by students for use in course selection, and by administrators for use in personnel decisions (p. 707), he also states that "studies of the usefulness of student ratings are both infrequent, and often anecdotal" (Marsh, 1987, p. 349)—or at least based more on survey data than on experimental data. The studies of short-term feedback reviewed by Cohen (1980) are the exception. Receiving feedback about student ratings administered during the first half of the term is **positively** related to improving college teaching as measured by student ratings administered at the end of the term. Based upon a meta-analysis of 22 studies, the student ratings at the end of the term vary with the kind of feedback:

During term	Student ratings at end of term
Instructors with **no** student rating feedback =	50th %ile
Instructors with **only** student rating feedback =	58th %ile
Instructors with **student rating feedback AND consultation** =	74th %ile

Conclusion

There are probably more studies of student ratings than of all of the other data used to evaluate college teaching combined. Although one can find individual studies that support almost any conclusion, for a number of variables there are enough studies to discern trends. In general, student ratings tend to be statistically reliable, valid, and relatively free from bias, probably more so than any other data used for faculty evaluation.

Nevertheless, student ratings are only one source of data about teaching and must be used in combination with multiple sources of data if one wishes to make a judgment of all of the components of college teaching. Further, student ratings are data that must be interpreted. We should not confuse a source of data with the evaluators who use the data to make a judgment.

References and Further Readings

In addition to the references cited in the text, I have tried to include all of the major reviews of the literature (indicated by an *) since 1970. For the general reader, I recommend Marsh's 1984 review. His 1987 review is essentially an elaboration and minor update of the 1984 review, more of interest to researchers.) For those interested in specific topics, I suggest the in-depth reviews published by Feldman. All of the major reviews will provide references to specific studies for the interested reader.

* Aleamoni, L. M. (1981). Student ratings of instruction. In J. Millman (Ed.), *Handbook of teacher evaluation* (pp. 110-145). Beverly Hills, CA: Sage.

* Aubrecht, J. D. (1979). *Are student ratings of teacher effectiveness valid?* (IDEA Paper No. 2). Manhattan: Kansas State University, Center for Faculty Evaluation and Development. (ERIC Document Reproduction Service No. ED 202 410)

* Aubrecht, J. D. (1981). *Reliability, validity and generalizability of student ratings of instruction.* (IDEA Paper No. 6). Manhattan: Kansas State University, Center for Faculty Evaluation and Development. (ERIC Document Reproduction Service No. ED 213 296)

Basow, S. A., & Distenfeld, M. S. (1985). Teacher expressiveness: More important for male teachers than female teachers? *Journal of Educational Psychology, 77*, 45–52.

Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology, 79*, 308–314.

Bausell, R. B., Schwartz, S., & Purohit, A. (1975). An examination of the conditions under which various student rating parameters replicate across time. *Journal of Educational Measurement, 12*, 273–280.

* Benton, S. E. (1982). *Rating college teachers: Criterion validity studies of student evaluation-of-instruction instruments.* AAHE-ERIC Higher Education Research Report, No. 1. Washington, DC: American Association for Higher Education.

* Braskamp, L. A., Brandenberg, D. C., & Ory, J. C. (1984). *Evaluating teaching effectiveness: A practical guide.* Beverly Hills, CA: Sage Publications.

Cashin, W. E. (1985, March). *Student ratings: The need for comparative data.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 261 098)

Cashin, W. E., & Clegg, V. L. (1987, April). Are student ratings of different academic fields different? Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED 286 935)

Cashin, W. E., Noma, A., & Hanna, G. S. (1987). *IDEA technical report no. 6: Comparative data by academic field.* Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E., & Perrin, B. M. (1978). *IDEA technical report no. 4: Description of standard form data base.* Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E., & Slawson, H. M. (1977). *IDEA technical report no. 2: Description of data base 1976–1977.* Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

* Centra, J. A. (1979). *Determining faculty effectiveness: Assessing teaching, research, and service for personnel decisions and improvement.* San Francisco, CA: Jossey-Bass.

* Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: A meta-analysis of findings. *Research in Higher Education, 13*, 321–341.

* Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*, 281–309.

Cohen, P. A. (1986, April). An updated and expanded meta-analysis of multisection student rating validity studies. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

* Costin, F., Greenough, W. T., & Menges, R. J. (1971). Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research, 8*, 511–535.

* Feldman, K. A. (1976a). Grades and college students' evaluations of their courses and teachers. *Research in Higher Education, 4*, 69–111.

* Feldman, K. A. (1976b). The superior college teacher from the students' view. *Research in Higher Education, 5*, 243–288.

* Feldman, K. A. (1977). Consistency and variability among college students in rating their teachers and courses: A review and analysis. *Research in Higher Education, 6*, 233–274.

* Feldman, K. A. (1978). Course characteristics and college students' ratings of their teachers, what we know and what we don't. *Research in Higher Education, 9*, 199–242.

* Feldman, K. A. (1979). The significance of circumstances for college students' ratings of their teachers and courses. *Research in Higher Education, 10*, 149–172.

* Feldman, K. A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education, 18*, 3–124.

* Feldman, K. A. (1984). Class size and college students' evaluations of teachers and courses: A closer look. *Research in Higher Education, 21*, 45–116.

* Feldman, K. A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education, 24*, 129–213.

* Feldman, K. A. (1987). Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: A review and exploration. *Research in Higher Education, 26*, 227–298.

* Feldman, K. A. (in press). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? *Research in Higher Education.*

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimation of the teacher and course components. *Journal of Educational Measurement, 15*, 1–13.

Hogan, T. P. (1973). Similarity of student ratings across instructors, courses, and time. *Research in Higher Education, 1*, 149–154.

Howard, G. S., & Maxwell, S. E. (1980). The correlation between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology, 72*, 810–820.

Howard, G. S., & Maxwell, S. E. (1982). Do grades contaminate student evaluations of instruction? *Research in Higher Education, 16*, 175–188.

* Kulik, J. A., & Kulik, C. C. (1974). Student ratings of instruction. *Teaching of Psychology, 1*, 51-57.

* Kulik, J. A., & McKeachie, W. J. (1975). The evaluation of teachers in higher education. In F. N. Kerlinger (Ed.), *Review of research in education* (Vol. 3, pp. 210-240). Itasca, IL: F. E. Peacock.

Marsh, H. W. (1982). The use of path analysis to estimate teacher and course effects in student rating's of instructional effectiveness. *Applied Psychological Measurement, 6*, 47-59.

* Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*, 707-754.

* Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 11*, 253-388.

Marsh, H. W., & Overall, J. U. (1979). Long-term stability of students' evaluations: A note on Feldman's consistency and variability among college students in rating their teachers and courses. *Research in Higher Education, 10*, 139-147.

Marsh, H. W., Overall, J. U., & Kesler, S. P. (1979). Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluation by their students. *Journal of Educational Psychology, 71*, 149-160.

* Marsh, H. W., & Ware, J. E. (1982). Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: New interpretations of the Dr. Fox effect. *Journal of Educational Psychology, 74*, 126-134.

* McKeachie, W. J. (1979). Student ratings of faculty: A reprise. *Academe, 65*, 384-397.

* Menges, R. J. (1973). The new reporters: Students rate instruction. In C. R. Pace (Ed.), *Evaluating learning and teaching. New directions for higher education, No. 4* (pp. 59-75). San Francisco, CA: Jossey-Bass.

* Miller, R. I. (1987). *Evaluating faculty for promotion and tenure*. San Francisco, CA: Jossey-Bass.

* Murray, H. G. (1980). *Evaluating university teaching: A review of research*. Toronto, Canada: Ontario Confederation of University Faculty Association.

Naftulin, D. H., Ware, J. E., & Donnelly, F. A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education, 48*, 630-635.

Overall, J. U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology, 72*, 321-325.

* Seldin, P. (1980). *Successful faculty evaluation programs: A practical guide to improve faculty performance and promotion/tenure decisions*. Crugers, NY: Coventry Press.

IDEA Papers may be ordered from the Center. Individual copies are $1.00. A complete set of IDEA Papers may be ordered for $5.00. Bulk orders: 25-99 copies, 25 cents per copy; 100 or more copies are 20 cents per copy. Orders of less than $25.00 must be prepaid.